图书情报工作
第66卷 第12期 2022年6月

突发传染病情境下社会化问答平台用户角色形成及转变*
——以知乎平台为例

陈苗苗[1] 安璐[1,2]
[1] 武汉大学信息管理学院 武汉 430072 [2] 武汉大学信息资源研究中心 武汉 430072

摘 要：[目的/意义]探究突发传染病情境下问答平台用户角色分类方法、角色形成关键因素及转变特点和差异。[方法/过程]收集问答平台 Covid-19 疫情数据相关数据共计 702 927 条，从参与程度和价值维度识别用户角色，基于信息人因子、信息因子和信息环境因子识别社区用户角色形成的影响因素，结合多分类模型和 SHapley Additive exPlanations（SHAP）模型分析影响不同角色形成的关键因素，利用 FP-growth 关联规则算法挖掘不同角色转变下的行为模式和主题特点。[结果/结论]研究结果表明用户倾向于维持角色不变且转变方向以积极型和潜水型为主，信息量是不同角色形成的关键因素，不同转变阶段的用户角色转变特征变化程度及所有转变阶段的用户角色转变行为具有显著差异。

关键词：用户角色 知乎问答平台 角色转变 影响因素 Covid-19 突发传染病
分类号：G252
DOI：10.13266/j.issn.0252-3116.2022.12.007

1 引言

2020 年初新冠肺炎疫情的突然暴发，不仅危害公众健康，也让社交媒体平台得到了空前的关注与利用，隔离状态下的公众使用社交媒体等互联网工具进行全面的信息搜索与及时的信息沟通，但突发公共卫生事件的不确定性常常引起公众恐慌与焦虑，进而使其转向在线问答平台寻求帮助和信息满足。面对疫情，习近平总书记强调"要把控好整体舆论，努力营造良好舆论环境。要加强网络媒体管控，推动落实主体责任、主管责任、监管责任，对借机造谣滋事的，要依法打击处理。"[1]。在此背景下，用户成为社会舆论参与的重要主体，不论是社区管理者还是政府部门都应加强对突发事件情境下用户的了解，包括其扮演的角色、行为特点和转变差异等，这将帮助有关部门制订针对性的用户政策，促进良好的信息交流环境。

知乎平台是社会化问答平台的一种，也是当前国内最大的知识问答平台，2021 年月活跃用户高达 1 亿人次，在该平台中，用户可以浏览、搜索、关注、提问、回答以达到满足自己的信息需求或者信息交流的目的。但现有的大多数关于用户角色的研究主要以微博、推特等社交平台的社区为主，不同的平台因其设计模式、定位等方面的不同，在用户角色、行为特征及具体的研究方案设计上有所差异，因此，对知乎平台在新冠肺炎疫情场景下的用户群体进行分析，也可以补充相关的研究。

针对虚拟社区用户角色的研究，主要集中在从用户行为或用户网络结构进行用户角色识别和分类上[2]，归纳总结社区用户的行为模式，包括参与行为、互动行为等，发现了一些经典的角色，如潜水者、贡献者等，而由于社会化问答社区提问和回答行为的特别性，其用户角色通常分为潜水者、提问者和回答者[3]，贡献程度依次递增，也因此吸引了学者们研究用户角色变化动因，研究方法则主要使用问卷调查和访谈法。总体上而言，对虚拟社区用户角色的研究倾向于静态层面的用户角色识别和分类；其次，社会化问答平台的角色细分通常划分为提问者和回答者，未能充分反映社会化问答平台的用户行为模式；最后，学者们主要关

* 本文系国家自然科学基金面上项目"危机情境下网络信息传播失序识别与干预方法研究"（项目编号：72174153）、国家自然科学基金创新研究群体项目"信息资源管理"（项目编号：71921002）和国家自然科学基金重大课题"国家安全大数据综合信息集成与分析方法"（项目编号：71790612）研究成果之一。
作者简介：陈苗苗，博士研究生；安璐，数据管理与知识服务研究室主任，教授，博士，博士生导师，通信作者，E-mail：anlu97@163.com。
收稿日期：2021-12-30 修回日期：2022-04-06 本文起止页码：68-81 本文责任编辑：易飞

注角色演变动因,鲜少探究社区用户角色形成的关键要素。基于此,本文的研究问题在于如何从用户行为模式特点识别社区用户角色? 对于各类用户角色而言,其角色形成的关键要素是什么? 各类用户角色发生转变的主要行为和问答的主题特点是什么? 本研究以新型冠状病毒(英文简称为 Covid-19)疫情为例,收集知乎问答平台新型冠状病毒话题下问答数据和用户数据累计 702 927 条,从问答行为和问答质量重新划分、解释社会化问答平台用户角色,结合信息生态理论和知乎平台特点,识别社区用户角色形成的影响因素,并利用 SHAP 模型探究每种因素对不同用户角色形成的影响,最后使用关联规则方法分析基于信息人因子和信息因子及主题的用户角色转变规律,揭示突发公共卫生事件情境下用户角色演化行为和问答主题的特点。

2 文献回顾与理论基础

2.1 虚拟社区用户角色研究

虚拟社区用户角色研究主要围绕社交网络结构特征和用户行为特征对用户角色进行识别和分类。前者将用户视为交互行为网络中的一员,其交互节点的网络拓扑结构如度中心度、紧密度中心度等决定了用户角色,一般划分为信息生成者、信息驱动者和信息桥接者[4],也有的根据入度和出度的情况将问答平台用户角色划分为学习成长型、乐于助人型、善于思考型、默默学习型和偏好不定型[5]。后者的角色识别通常与用户参与社区的行为相关,如早期 J. Hagel 等按照参与程度将社区用户角色定性地划分为浏览者、潜水者、贡献者和购买者[6],在后续研究中,多使用更加科学的方法如统计分析、聚类方法识别用户角色,包括意见领袖、专家等[7],如 J. Villodre 从社交媒体和应急管理的视角,根据用户发布推文的数量重新定义了影响者、传播者和普通用户[8]。

近年来,部分研究开始关注角色转变。一方面发现角色转移方式,如 J. Preece 等提出用户通过线性或非线性的方式从一个角色转移到另一个角色[9],但是缺乏实验支持;C. Fu 提出用时间感知角色模型来有效跟踪用户角色的演变[10],将用户角色简单抽象为提问者和回答者并主要关注模型构建;A. Bartal 等基于动态网络提出时间角色归属频率模型,识别出了有影响力的成员[11]。另一方面,关注角色变化的动因,尤其是单一角色转变的动因,主要通过问卷调研的方式来挖掘,如赵欣等发现专业知识与互惠规范是导致知识搜寻者向知识贡献者转变的内在因素[12],G. Zeng 等发现潜水者向知识贡献者转变的关键是自我效能、用户信任、共同愿景和社区忠诚度等[13]。

2.2 信息生态相关理论

信息生态学引入自然生态的概念,从生态学视角探究与信息系统、信息自组织等相关的问题,是一门研究信息规律的科学[14－15],强调信息生态因子间的相互和谐。1999 年 B. A. Nardi 等在 *Information ecologies: using technology with heart* 一书中率先提出信息系统的概念并认为信息生态是“由人、实践、价值和技术在特定环境中所组成的系统”[16],将人、实践、价值和技术称为信息生态因子。但学者们对信息生态因子有不同的见解,逐渐形成了二要素理论(信息人和信息生态环境)[17]、三要素理论(信息、信息人和信息环境)[188]和四要素理论(信息、信息人、信息技术和信息环境)[19]三种学说。信息、信息人和信息环境这三个要素被认为是信息生态系统中最为核心的要素,信息技术是从信息环境中割离出来的,考虑到所有进入社区的用户在信息技术的接触上没有差别,在本文中主要应用三因素理论构建社区用户角色形成的影响要素。

2.3 生命周期理论

在突发事件生命周期的研究中,经典理论包括 S. Fink 提出的潜伏期、暴发期、延续期和痊愈期四阶段模型[20]和 B. T. Burkholder 等提出的事前、事中和事后紧急事件管理三阶段模型[21],以这些理论为基础,安璐等通过划分生命周期比较各突发事件严重性指标在不同生命阶段的区别,并构建了预警机制[22]。刘冰等依据 S. Fink 的生命周期理论构建重大突发公共卫生事件风险研判与决策模型[23]。姜金贵等结合主题和情绪走向将生命周期划分为形成期、高潮期、波动期和消散期[24],在一些突发事件中,舆情生命周期可能会以单峰型、双峰型及多峰型多种方式演进[25]。在本研究中,拟根据突发公共卫生事件的特点,将整个事件的演化阶段划分为潜伏期、暴发期、第一次衰退阶段、波动期、第二次衰退阶段和平息阶段,其中,波动期是指事件发展到一定高峰后,会经历一段时间的沉寂,而随着新信息的刺激,又出现新高潮并且呈现波峰和波谷交替摆动的形态。

综上所述,在用户角色研究上,已有研究主要集中于角色识别分类的静态分析,而角色转变研究仅关注某一类角色转变,如知识寻求者向知识贡献者的转变,缺乏对角色形成的关键因素分析以及从整体的角度考虑全部角色转变的动态分析和差异分析,也较少关注

问答平台上的角色情况，而不同社交媒体平台的用户角色及行为特点是具有一定差异的。因此，本文拟针对问答平台，识别适用于问答平台的用户角色，并基于信息生态理论和生命周期理论分析角色形成的关键因素及不同角色转变的动态特点和差异。

3 研究方法

3.1 数据收集与预处理

本文收集知乎问答平台“新型冠状病毒”话题下 2019 年 12 月 30 日 – 2020 年 5 月 31 日所有问题及回答，经过去重后累计获取相关问题、回答及文章共计 466 274 条，其中问题 15，401 条、文章 976 条、回答 449 897条，累计涉及 236 653 名实名用户和 42 063 名匿名用户，根据用户 id 进一步收集了这 236 653 名用户的详细信息，获得初始的问答及用户数据共计 702 927条。然后人工去除不相关问题及回答、文章以及已注销用户和匿名用户的提问及回答数据，最终得到关于新型冠状病毒话题下的有效提问和回答数据共计407 247条，涉及 236 304 名用户。

3.2 问答平台用户角色分类

在问答平台中，用户的提问和回答行为被认为是最有价值的部分，本文借鉴 J. Hagel 等从参与程度和价值两个角度定性地划分社区用户角色的思想[6]，将问答平台用户角色重新解释和划分为潜水型、积极型、需求型和知识型，试图从量化的角度识别适用于问答平台的用户角色类别。其中，潜水型对应于 J. Hagel 等所提出的潜水者，该角色对社区所贡献的信息很少；积极型对应于其所提出的贡献者，通常用户十分活跃；需求型和知识型对应于其所提出的购买者，并根据问答平台的特点细化为需求型和知识型。不同于 Hagel 等设定参与程度和价值维度仅与用户行为相关，在本文中，参与程度是指各个生命周期内用户提问和回答的频率，而价值则是指各个生命周期内用户提问和回答的质量。

问答平台中用户的答案质量常使用获赞数衡量[26]，在危机情境下，也有部分学者使用获赞数代表对信息的认可程度或者采纳程度[27]。由于知乎平台无法获取关于问题回答的“喜欢”“收藏”这两个指标，因此，本文采用获赞数来衡量答案质量。关于问题质量的研究较为匮乏，李胜利等根据 CSDN 和 Stack Overflow 两个社区的特点以问题得分和问题得分分布情况评估问题的质量[28]，类似地，根据知乎平台特点，可使用问题的获赞数量、问题关注数量和问题浏览数量来衡量问题的质量，问题的获赞数量表示对问题的认可程度，问题关注数量表示问题的代表程度，即代表具有相似问题需求的人的程度，问题浏览数量代表着问题的吸引力，换言之，一个高质量的问题应能获得社区其他用户的认可、代表大多数人的需求并吸引更多的注意力。因此，在用户角色分类价值维度上，用户的答案质量的评估计算如公式(1)所示，$Answer_{count}$ 表示用户所有回答的数量，$like_i$ 表示用户第 i 个回答的获赞数量：

$$Answer_{quality} = \frac{\sum_i^{Answer_{count}} like_i}{Answer_{count}} \qquad \text{公式（1）}$$

用户问题质量的评估计算如公式(2)所示，$Question_{count}$ 表示用户所有提问的数量，$like_i$ 表示用户第 i 个问题被认为是好问题的数量，$follow_i$ 表示用户第 i 个问题被关注的数量，$browse_i$ 表示用户第 i 个问题被浏览的数量，w_1、w_2、w_3 分别为这三个指标的权重：

$$Question_{quality} = \frac{\sum_i^{Question_{count}} w_1 * like_i + w_2 * follow_i + w_3 * browse_i}{Question_{count}} \qquad \text{公式（2）}$$

本文组合熵权法和变异系数法对变量权重赋值，前者依据指标的变异程度反映其信息量的大小来确定权重，后者通过衡量指标观测值变动程度确定权重，二者的结合可以使赋权结果更加准确。最终用于计算问题质量的组合权重如公式(3)所示，α 表示熵权法占组合权重的比例，$W_{E\text{-}Wj}$表示由熵权法计算得到的第 j 个指标的权重；$1-\alpha$ 表示变异系数法占组合权重的比例，$W_{C\text{-}Vj}$表示由变异系数法得到的第 j 个指标的权重。通常 α 系数取值为 0.5[29]：

$$w_j = \alpha * W_{E\text{-}Wj} + (1-\alpha) W_{C\text{-}Vj} \qquad j = 1,2,3 \qquad \text{公式（3）}$$

根据用户的两种行为和质量，针对每一生命周期阶段里的用户，知乎问答平台用户角色分类划分原则如下：

（1）潜水型用户。当用户仅提问时，$Question_{quality} <= AVG(\sum_1^n Question_{quality})$ 且 $Question_{count} <= AVG(\sum_1^n Question_{count})$，即用户的提问质量和提问数量均低于均值的时候，用户为潜水型用户，n 表示有提问行为的用户数量；或者当用户仅回答时，$Answer_{quality} <= AVG(\sum_1^m Answer_{quality})$ & $Answer_{count} <= AVG(\sum_1^m Answer_{count})$，即用户的回答质量和回答数量均低于均值的时候，用户亦为潜水型用户，m 表示有回答行为的用户数量。本文结合发帖频率和发帖质量定义潜水型用户为不积极参与社区且用户发帖价值较低的群体，在危机应对期间，这类

用户参与了社区但却无法对社区做出较大的贡献。

（2）积极型用户。当用户仅提问时，$Question_{quality} <= AVG(\sum_1^n Question_{quality})$ 且 $Question_{count} > AVG(\sum_1^n Question_{count})$，即用户提问质量低于均值且提问频率高于均值的时候，用户为积极型用户；或者当用户仅回答时，$Answer_{quality} <= AVG(\sum_1^m Answer_{quality})$ & $Answer_{count} > AVG(\sum_1^m Answer_{count})$，即用户回答质量低于均值且回答频率高于均值的时候，用户亦为积极型用户，其中 n 和 m 的定义与前文一致。积极型用户具有较高的活跃水平，但是他们所产出的价值并不高，这类用户往往具有较高的社交需求[30]。

（3）需求型用户和知识型用户。需求型用户具有较高的信息需求，即提问质量高。知识型用户产生具有价值的信息，即回答质量高。需求型用户和知识型用户在社区中的共性是其问答质量较高，但在社区平台中，有的用户既产生提问行为又产生回答行为，为了区分该用户偏向需求型还是知识型，在得到各个生命周期阶段所有用户提问数量、回答数量、提问质量和回答质量后，使用最大最小标准化方法将数据进行标准化得到 $Question^*_{count}$、$Answer^*_{count}$、$Question^*_{quality}$、$Answer^*_{quality}$，以消除各指标的取值大小对角色划分的影响。借鉴信息论中对信息量的描述[31]，即事件发生概率与信息量呈现正相关，其事件 i 的信息量计算公式为 $H = -\log_2 P_i$，类比于该计算公式，本文区分需求型用户和知识型用户的计算公式如（4）所示。需要注意的是标准化后的数量和质量相乘不再具有原始意义，因为已消除量纲影响。

$$\text{s.t.}\begin{cases} R_{demand} = -\log_2 \dfrac{1}{Question^*_{count} * Question^*_{quality} + 1} \\ R_{knowledge} = -\log_2 \dfrac{1}{Answer^*_{count} * Answer^*_{quality} + 1} \\ user_{role} = \text{知识型用户} \quad R_{knowledge} \geq R_{demand} \\ user_{role} = \text{需求型用户} \quad R_{knowledge} < R_{demand} \end{cases}$$

公式（4）

3.3 社区用户角色形成的影响因素识别

根据信息生态理论，本文从信息人因子、信息因子和信息环境因子分析社区用户角色形成的影响因素，如表 1 所示：

表 1 信息生态视角下的影响因素

生态因子	信息生态因子表征		特征值
信息人因子	自然属性	性别	男/女/未填写
		地域	北京/上海/广州......
		所在行业	电子商务/公共服务/互联网......
		创作等级	[0-10]
		用户类别	个人/组织
		是否认证	是/否
		是否为优秀回答者	是/否
		粉丝数	用户的粉丝数
		关注数	用户的关注数
	特征属性	影响力	粉丝数与（粉丝数+关注数+1）的比值
		绝对角色	提问者/回答者/既是提问者又是回答者
		上一阶段用户角色	潜水型/积极型/知识型/需求型
		兴趣程度	用户话题分布同该环境下的话题相似性
信息因子	文本属性	提问标题平均长度	用户提问标题总长度与提问数量的比值
		提问描述平均长度	用户提问描述总长度与提问数量的比值
		回答文本平均长度	用户回答文本总长度与回答数量的比值
		文本信息量	问答文本中所有词的 TF-IDF 之和
	情感属性	提问标题情感倾向值	提问标题文本的情感积极倾向值的和与提问数量的比值
		提问描述情感倾向值	提问描述文本的情感积极倾向值的和与提问数量的比值
		回答文本情感倾向值	回答描述文本的情感积极倾向值的和与回答数量的比值
	时间属性	平均问答时间间隔	用户所有回答与问题的时间间隔同所有提问数量和回答数量的比值
		时间分布	分为深夜（00:00-06:00]、清晨（6:00-8:30]、上午（8:30-12:00]、中午（12:00-14:00]、下午（14:00-18:00]、晚上（18:00-24:00]，以是/否在该时间段有文本发布表示
	主题属性	主题分布	单个用户所有文本信息主题在所有用户全部文本信息上的主题分布，以是/否包含某个话题表示
		主题丰富性	单个用户所有文本包含的主题总数
信息环境因子	环境属性	平台介入度	问题回答折叠数量和（问题回答数量+1）的比值
		信息讨论度	问题回答数量
			回答评论数量
		演化阶段	潜伏期、暴发期、第一次衰退阶段、波动期、第二次衰退阶段、平息阶段

3.3.1 信息人因子

信息人因子通常使用自然属性表征，即用户的年龄、性别等社交媒体自然属性[32]，本文中还提出信息人因子包含特征属性，即为用户的非自然属性但与用户本身相关的属性。根据知乎平台特点，信息人自然属性如表1所示，其中，地域重新划分取值36个特征值，包括中国34个省级行政区及海外和其他地域；所在行业使用知乎提供的行业信息，累计涉及112个特征值；创作等级从0级到10级，共11个特征值。

信息人特征属性表征如表1所示，其中，绝对角色为用户在某一阶段的提问行为和回答行为，如果用户只提问，则为提问者；如果只回答，则为回答者；如果既提问又回答，则既是提问者又是回答者。上一阶段用户角色为用户在上一生命阶段的角色，由于用户并不总是存在于社区中，故另外设置两个虚拟角色：用户进入和用户退出，如果在某一生命周期，用户第一次进入该社区，则其上一阶段角色为用户进入，如果用户并不是第一次进入该社区且上一生命周期未发过任何帖子，则上一阶段用户角色为用户退出。用户的兴趣程度计算如公式(5)所示，表示用户所有文本主题分布概率与该生命周期阶段文本主题分布概率的相似性，主题分布概率由全部用户 n 在所有主题分布概率之和的平均表示，本文使用 JS 散度评估二者的相似性，其取值范围为[0,1]，值越小表明用户越感兴趣。

$$Interest_degree = JS_divergence\left(P\{t_1, t_2, \cdots, t_r\}, P\left\{\frac{\sum_1^n t1}{n}, \frac{\sum_1^n t2}{n}, \cdots \frac{\sum_1^n tr}{n}\right\}\right) \quad \text{公式 (5)}$$

3.3.2 信息因子

信息因子可细化为信息时效性、信息有用性等特征[33]，考虑到信息因子描述信息的基本特点，本文将信息因子分为文本属性、情感属性和时间属性。文本属性除了统计提问标题、提问描述和回答文本的平均长度外，还使用 TF-IDF 计算用户所有帖子的总体信息量，即利用去除停用词后的用户发布的所有问答中词的 TF-IDF 值的加和代表用户在某一生命周期发表的所有帖子的信息量。在用户情感属性的计算上，本文借助百度情感分析模型 Senta 中的 Bi-LSTM 预测帖子情感[34]，首先对用户所发的帖子进行分句处理，并使用积极情感倾向概率作为文本的情感倾向值，其计算如公式(6)所示，表示某个用户发布的全部帖子的情感倾向值，m 表示用户累计发的提问或者回答的帖子数量，n 表示提问帖子或者回答帖子的句子数量，$P(Sentence_{ji})$ 表示用户第 j 个帖子的第 i 句话的积极情感倾向。

$$Emotion_{user} = \frac{\sum_{j=1}^{m} \sum_{i=1}^{n} P(Sentence_{ji})}{n * m} \quad \text{公式 (6)}$$

时间属性包括时间分布和动态时间分布间隔，其中时间分布的划分方式如表1所示，每个用户在某个生命周期发布的帖子都将离散地分布在表中的6个时间段，并以是/否的形式表示。关于动态时间分布间隔，将用户提问行为视为问题首发，因此时间间隔为0，用户回答行为时间间隔为用户回答时间与其回答的问题的时间间隔，最终动态时间间隔为用户所有回答时间同问题的时间间隔和所有回答数量和提问数量之和的比值。

主题属性包括主题分布和主题丰富性。本文使用基于 Bert 的文本聚类工具 Bertopic 识别用户发布信息主题，Bertopic 利用 transformers 和 c-TF-IDF 创建密集的集群，得到的主题易于解释[35]。主题丰富性解释见表1。

3.3.3 信息环境因子

关于信息环境因子，在社交媒体中其外部环境考虑了转发、评论等属性[30]。本文的信息环境因子考虑了知乎平台介入度、信息讨论度和演化阶段，具体衡量见表1。

3.4 用户角色形成的影响因素分析

为了探索用户角色形成的关键因素，对表1影响因素为无序分类变量的，即地域和所在行业进行独热编码处理，将用户的性别、用户类别、是否认证、是否为优秀回答者等进行序列编码处理。然后将用户在该阶段的用户角色作为因变量，通过建立多分类模型，识别最佳的分类器，并将分类器作为 SHAP(SHapley Additive exPlanations)输入进行训练，并绘制不同角色的特征重要性排序图。SHAP 是基于博弈论衡量模型的特征重要性，具有可解释性[36]。在本文中，构建的多分类器包括线性回归分类器(LR)、K 近邻分类器(KNeighborsClassifier)、神经网络分类器(MLPClassifier)、决策树分类器(DecisionTreeClassifier)、随机森林分类器、LightGBM 分类器(LGBMClassifier)、CatBoost 分类器和 XGBoost 分类器。

在建立多分类模型时，考虑到多分类样本存在不均衡问题，本文通过过采样方法 Borderline-Smote 算法对训练集上的数据进行数据均衡操作，该算法能够更准确地学习每个类的边界，从而改善样本的类别分

布[37]。为识别最佳分类器,通过结合 Borderline-Smote 算法和十折交叉验证对不同多分类器进行评估,即在每次评估中都应用 Borderline-Smote 算法平衡训练集上的数据,并利用测试集预测效果评估多分类器结果。

3.5 用户角色转变分析

为了挖掘用户角色转变的规律,本文计算角色转变的支持度和置信度,以及不同生命周期阶段转换期间不同用户角色转变的主要指标变化程度,如公式(7)所示,表示某个转换阶段(A 阶段 - B 阶段)用户角色 U1 向用户角色 U2 转变的某一影响因素 i 的变化程度。其中,U2(i)为 B 阶段用户角色 U2 影响属性 i 的均值,U1(i)为 A 阶段用户角色 U1 影响属性 i 的均值,为避免分母为 0,将分母数值加 1。

$$\text{DegreeChange}_{(U1,U2)_i} = \frac{U2(i) - U1(i)}{U1(i) + 1}$$

公式 (7)

此外,本文使用 FP-growth 关联规则对不同角色转变条件下的信息人因子和信息因子、话题和话题因子进行关联分析,提取其潜在联系,从而发现不同角色转变的异同点。FP-growth 采取分治策略加速了关联规则挖掘过程[38],克服了 Apriori 算法效率低的问题。

在进行关联规则分析前,需要对数据进行类别处理,按高于平均值和低于平均值的二分类法划分粉丝数、关注数、影响力、提问标题平均长度、提问描述平均长度、回答文本平均长度、文本信息量、动态时间分布间隔这些特征;采取四分法将兴趣程度划分为 4 个类别,即按值[0,0.25]、(0.25,0.5]、(0.5,0.75]、(0.75,1]划分为非常感兴趣、很有兴趣、一般感兴趣和不太感兴趣;采用 Senta 官方分类标准,将各情感倾向值按[0,0.45]、(0.45,0.55)、[0.55,1]划分为消极、中立和积极。然后通过调节最小支持度和最小置信度并基于支持度、置信度和提升度找到各角色转变条件下的最相关和最有效的关联规则,选择最大提升度下的最大支持度的规则作为最终的关联规则,提升度越大说明两者之间越相关。为了简化结果形式,本文设计了一些表示规则,以“低影响力,[C(个人|不是优秀回答者)]”为例,[ ]表示其含有的内容可有可无,C(A|B)表示{A、B}除空集外的所有子集,因此,该例子可以形成“低影响力”“低影响力、个人”“低影响力、不是优秀回答者”“低影响力、个人、不是优秀回答者”共 4 种子规则,如果规则中有减号,表示不包含减号后的数据集的任意子集。对于规则前置项和后置项交换后的支持度、置信度和提升度一致,直接合并写成“A|B”的形式。

4 实验与结果分析

4.1 数据生命周期划分

依据网络信息空间传播特点,通过识别社交媒体信息数量变化的拐点划分生命周期,考虑到在该话题下匿名用户和已注销用户发布的帖子与该话题也密切相关,因而在划分生命周期时,仅去除无关数据,最终得到如表 2 所示的生命周期阶段,共计 6 个阶段。

表 2 生命周期划分

生命周期	时间区间	帖子数量/个	关键事件
潜伏期	2019.12.30-2020.1.20	268	1 月 20 日晚钟南山医生确认新冠病毒具有“人传人”现象,进入暴发期
暴发期	2020.1.21-2020.2.1	50 334	2 月 1 日武汉火神山医院即将建成,并于 2 月 2 日正式交付,进入第一次衰退阶段
第一次衰退阶段	2020.2.2-2020.3.11	117 959	3 月 11 日,世界卫生组织宣布新冠肺炎疫情为全球大流行,进入波动期
波动期	2020.3.12-2020.4.4	183 351	4 月 4 日全国哀悼;全球单日新增确诊新冠病例超过 10 万例。进入第二次衰退期阶段
第二次衰退阶段	2020.4.5-2020.5.3	96 676	5 月 3 日,国家卫健委:全国现有确诊病例连续 11 天下降;新增无症状感染者 12 例,为通报以来最低。进入平息阶段
平息阶段	2020.5.3-2020.5.31	17 224	

4.2 问答平台用户角色分类和转变情况

根据 3.2 节对问答平台用户角色划分的方法,我们首先按照生命周期阶段识别出每个阶段的用户,最终得到所有生命周期的用户共计 289 259 名,潜水型用户 221 349 名,占比 76.52%;积极型用户 52 925 名,占比 18.3%;知识型用户 14 503 名,占比 5.01%;需求型用户 482 名,占比 0.17%。我们统计了每个生命周期阶段不同角色的性别、行业、地域、创作等级、粉丝数、关注数及影响力的特征,发现用户主要聚集在北京、上海和广东省这类经济发达地区。积极型用户平均粉丝数稳定在 2 000-5 000 之间,其创作等级、影响力伴随生命周期的发展而逐步增加。潜水型用户在疫情发展初期粉丝数在 3 000 左右,疫情中期粉丝数下降至 1 000以下,后期又略微回升,其关注数相对稳定,影响力水平在疫情发展中期一直低于 0.4。需求型用户较为特殊,疫情发展的初期和后期,知乎官方平台为了促

进讨论氛围,会自主提出问题,导致需求型用户粉丝数和关注数较高,但影响力水平在 0.5 左右、创作等级在 6 左右。知识型用户无论是粉丝数、影响力还是创作等级,在疫情各阶段基本上处于最高水平,影响力水平在 0.8 左右,创作等级在 6.5 以上。另外,在性别、地域和行业上,知识型用户信息较全,但潜水型用户在这些特征上缺失值最多,各类型用户角色主要聚集在互联网、临床医疗、高等教育等行业,互联网行业的用户在任何角色中始终占据一席之地,而疫情发展前期临床医疗行业的用户也是关注疫情的重点用户。

根据 3.5 节对用户角色转变的描述,各类角色转变的情况如表 3 所示,从表 3 中可以看到不同角色转变概率不同,用户维持积极型不变的概率是 49.78%,从积极型转向潜水型的概率是 43.88%,而从积极型转向需求型或者知识型的概率低很多,其他角色转变概率具体如表 3 所示。通常情况下,用户倾向于维持原有角色不变,但需求型用户会以更大的概率转向潜水型或者积极型,潜水型和积极型会以较高的概率相互转换,但是积极型更容易向潜水型转变,需求型会以 13.8% 的概率向知识型转换,但知识型向需求型转换的概率较低,另外如果发生转变用户也会倾向于向潜水型或者积极型转变,需求型向各个角色转变的概率都较高。

表 3 角色转变情况

前项	后项	支持度	置信度	前项	后项	支持度	置信度
积极型	积极型	0.189 01	0.497 78	需求型	积极型	0.000 89	0.317 07
积极型	潜水型	0.166 63	0.438 83	需求型	潜水型	0.000 91	0.325 20
积极型	需求型	0.000 77	0.002 04	需求型	需求型	0.000 61	0.219 51
积极型	知识型	0.023 29	0.061 35	需求型	知识型	0.000 39	0.138 21
潜水型	积极型	0.148 20	0.295 22	知识型	积极型	0.033 48	0.289 94
潜水型	潜水型	0.328 30	0.653 98	知识型	潜水型	0.033 44	0.289 54
潜水型	需求型	0.000 59	0.001 18	知识型	需求型	0.000 61	0.005 32
潜水型	知识型	0.024 91	0.049 62	知识型	知识型	0.047 95	0.415 21

根据 3.5 节,本文挖掘了不同生命周期阶段用户角色发生转变的特征变化。如表 4 所示,因文章篇幅限制,仅展示各阶段角色转换概率排名前二及正负指标变化程度最高的前 5 位。转换概率表示该角色转换数量占该转换阶段所有角色转换数的比例。在不同生命周期转变期间,主要的转变角色略有差异,但潜水型用户维持角色不变的概率最大,每一转换期间,用户角色维持不变和变化的主要变化指标及变化程度也是不同的。以潜水型角色维持不变为例,前两个转变时期要求用户角色粉丝数变化度是前一阶段的 0.68 倍时才能维持角色不变,而后续如第二次衰退阶段到平息阶段,用户角色粉丝数变化度要求为前一阶段的 1.95 倍时,才有可能维持潜水型角色不变。在潜伏期到暴发期阶段,知识型用户维持不变的要求是其所发布的信息内容的信息量变化度是前一阶段的 0.62 倍,时间上要在上午或者深夜发布,主题上要更加聚焦。在不同转变阶段,角色发生转变的要求也不一样,以潜水型向积极型转变为例,从暴发期到第一次衰退阶段,用户从潜水型向积极型转变,意味着粉丝数变化度是前一阶段的 1.79 倍,且提问描述平均长度变化度是前一段的 3.75 倍、提问标题平均长度变化度是前一阶段的 1.55倍,信息量变化度是前一阶段的 0.76 倍等。

4.3 用户角色分类模型的训练与评估

在影响因素主题分布衡量上,由于 Bertopic 训练花费成本较高,在经过 5 天训练后得到 1 584 个主题,相当一部分与主题相关的文档数量仅在 10 篇左右,为了确保每个主题都至少有 100 篇文档,结合主题相似性矩阵,最终确认主题数为 180,并通过主题相似性对现有主题进行了进一步的合并。

根据 3.4 章节的描述,使用 Borderline-Smote 和十折交叉验证对模型进行评估,模型评估采用精准率、召回率、F1 值的宏平均结果及正确率。从表 5 中可以看到 CatboostClassifier 和 LGBMClassifier 在多分类上的表现都有最优值,但 CatboostClassfier 在精准率、F1 值和正确率上都表现优于 LGBMClassifier,故而在利用 SHAP 模型挖掘影响因素的时候,使用 CatBoostClassifier 作为其输入。

4.4 社区用户角色形成的影响因素分析

为了探究不同变量对于不同用户角色形成的具体影响,绘制特征重要性排序图(见图 1)。图 1 中颜色深浅表示特征值的大小,颜色越偏向于浅灰色,特征值越小,否则,特征值越大。

表 4　各生命周期转换期间不同角色转变的指标变化情况(部分)

转换期间	转换角色	转换概率	主要变化指标(括号内的数值为变化程度)	
			正向变化	负向变化
潜伏期－暴发期	潜水型→潜水型	0.31	粉丝数(0.68);动态时间分布间隔(0.64);性别(0.31);主题丰富性(0.16);中午(0.15)	回答评论数量(－0.48);关注数(－0.4);晚上(－0.27);回答情感倾向值(－0.04);用户类别(－0.02)
	知识型→知识型	0.17	动态时间分布间隔(1.98);粉丝数(1.08);信息量(0.62);上午(0.47);深夜(0.29)	主题丰富度(－0.49);回答评论数量(－0.21);性别(－0.09);晚上(－0.09);兴趣程度(－0.04)
暴发期－第一次衰退阶段	潜水型→积极型	0.19	提问描述平均长度(3.78);粉丝数(1.79);提问标题平均长度(1.55);关注数(0.99);信息量(0.76)	主题丰富性(－0.14);兴趣程度(－0.14);回答情感倾向值(－0.05)
	潜水型→潜水型	0.37	粉丝数(0.68);关注数(0.5);回答平均文本长度(0.26);性别(0.21);影响力(0.14)	上午(－0.09);深夜(－0.04);回答情感倾向值(－0.03);中午(－0.01)
第一次衰退阶段－波动期	潜水型→积极型	0.21	提问描述平均长度(3.75);提问标题平均长度(1.22);粉丝数(0.88);信息量(0.69);关注数(0.68)	主题丰富性(－0.11);动态时间分布间隔(－0.11);回答情感倾向值(－0.04);回答评论数量(－0.01);兴趣程度(－0.01)
	潜水型→潜水型	0.32	关注数(0.42);粉丝数(0.30);性别(0.17);创作等级(0.12);影响力(0.08)	动态时间分布间隔(－0.08);回答评论数量(－0.08);下午(－0.03);回答情感倾向值(－0.03);主题丰富性(－0.02)
波动－第二次衰退阶段	积极型→积极型	0.23	粉丝数(1.12);关注数(0.34);动态时间分布间隔(0.18);回答平均文本长度(0.10);影响力(0.09)	提问描述平均长度(－0.6);提问标题平均长度(－0.50);问题回答数量(－0.12);主题丰富性(－0.06);提问标题情感倾向值(－0.03)
	潜水型→潜水型	0.31	粉丝数(0.89);关注数(0.40);动态时间分布间隔(0.13);性别(0.12);影响力(0.08)	回答情感倾向值(－0.03);晚上(－0.03);回答平均文本长度(－0.01)
第二次衰退阶段－平息阶段	积极型→潜水型	0.31	关注数(0.36);主题丰富性(0.19);动态时间分布间隔(0.16);粉丝数(0.14);影响力(0.08)	提问描述平均长度(－0.91);提问标题平均长度(－0.78);信息量(－0.34);问题回答数量(－0.25);下午(－0.15)
	潜水型→潜水型	0.34	粉丝数(1.95);关注数(0.57);性别(0.15);影响力(0.11);创作等级(0.1)	回答平均文本长度(－0.07);主题丰富性(－0.05);回答情感倾向值(－0.02);下午(－0.02)

表 5　用户角色多分类器评估结果

模型	精准率	召回率	F1 值	正确率
LR	0.594 2	0.320 1	0.323 0	0.550 8
KNeighborsClassifier	0.541 0	0.462 6	0.479 0	0.640 0
MLPClassifier	0.789 6	0.651 0	0.660 3	0.874 6
DecisionTreeClassifier	0.823 3	0.803 5	0.807 6	0.914 7
RandomForestClassifier	0.771 1	0.854 4	0.803 1	0.913 5
LGBMClassifier	0.873 7	**0.881 8**	0.869 4	0.957 7
CatBoostClassifier	**0.881 3**	0.868 9	**0.872 9**	**0.966 2**
XGBClassifier	0.873 3	0.883 4	0.870 6	0.957 3

从图 1 中可以得到以下结论：

(1)对于潜水型用户而言,文本信息量越低、回答文本平均长度越长、提问标题平均长度越短、提问标题情感倾向值越低、动态分布时间间隔越高、回答情感倾向值越高、上一阶段用户角色为用户进入或潜水型、粉丝数越低、行业和地域不确定、回答评论数量越低、问题回答数量越低、主题丰富性越低、演化阶段越远时,用户则越有可能是潜水型用户。

(2)对于积极型用户而言,文本信息量越高、回答文本平均长度越低、提问标题情感倾向值越高、动态分布时间间隔越高、回答情感倾向值越低、上一阶段用户角色非用户进入、粉丝数越低、行业和地域不确定、回答评论数量越低、问题回答数量越低、主题丰富性越低、演化阶段越远时,用户则越有可能是积极型用户。此外,积极型用户除了清晨外,其他任意时间段都爱发帖子,而潜水型用户不喜欢在下午、上午及晚上发帖。

(3)对于知识型用户而言,信息环境因子是其促进其形成的一个重要因素,当回答评论数量越高、回答平均文本长度越长、问题回答数量越低、粉丝数越高、信息量越高、提问标题平均文本长度越低、动态时间分布间隔越低、主题丰富性越高、影响力越高、提问标题情感倾向值越低、性别偏向于男、创作等级越高时,则越能促进用户成为知识型用户。

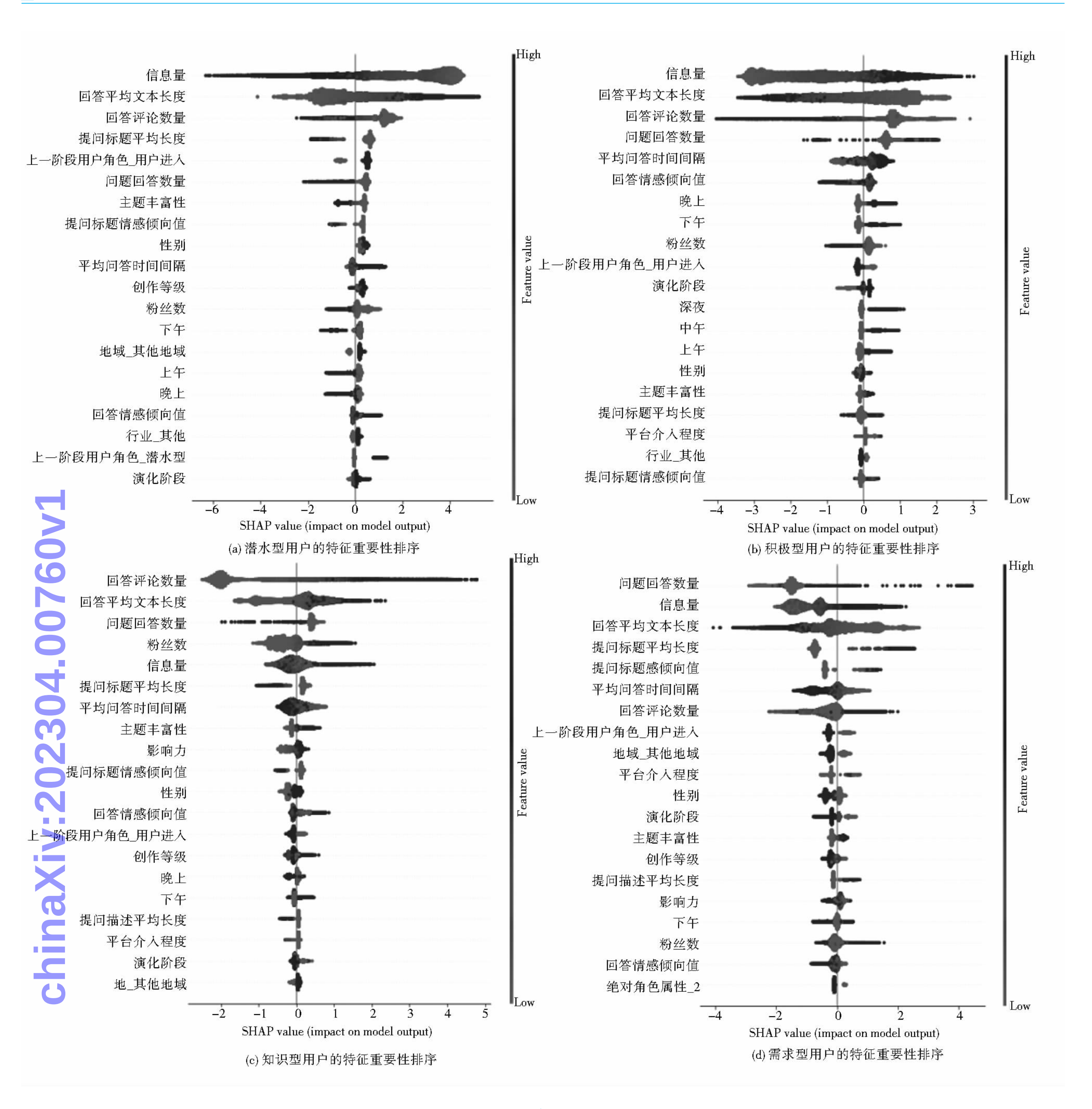


图 1　不同类型用户的特征重要性排序

（4）对于需求型用户而言，当问题回答数量越高、信息量越高、回答平均文本长度越短、提问标题平均长度越长、提问标题情感倾向值越高、动态时间分布间隔越短、回答评论数量越高、上一阶段用户角色为非用户进入、性别偏向于未填写、主题丰富性越高、创作等级越低、提问描述平均长度越高、绝对角色属性非回答者时，则越能促进用户成为需求型用户。

4.5　基于信息人因子和信息因子的用户角色转变分析

根据 3.5 章节，本文构建了如表 6 所示的不同角色转变条件下的关联规则。以用户角色从潜水型向需求型转变为例，当用户从潜水型向需求型变化的时候，该用户通常既是提问者又是回答者且对当前话题很有兴趣，并会以 78% 的概率在上午提出具有高信息量且提问标题字数较少的问题，这条规则揭示了用户角色从潜水型向需求型转变的信息人特征和信息行为特征。

本文根据表 6 的关联规则，总结了如下所示的角色转变规律：

（1）当用户从不同角色向潜水型角色转变的时候，其通常表现出回答文本长度较低或者文本内容包含的信息量较低的现象，但不同角色的转变具有一定

表6 各角色转变下信息人因子和信息因子的关联规则

角色转变	关联规则(信息人因子 = >信息因子)		支持度	置信度	提升度
潜水→潜水	低影响力,粉丝数低于均值,个人,性别_未填写	回答平均文本长度低于均值	0.22	0.80	1.13
	低影响力,关注数低于均值,其他地域,未认证,[粉丝数低于均值],[C(个人|回答者|不是优秀回答者)]	回答平均文本长度低于均值,低信息量	0.22	0.80	1.13
潜水→积极	低影响力,回答者,其他地域,[C(个人|粉丝数低于均值|不是优秀回答者)]	回答平均文本长度低于均值	0.23	0.88	1.20
潜水→知识	粉丝数高于均值	高信息量，回答平均文本长度高于均值	0.24	0.70	1.21
潜水→需求	既是提问者又是回答者,很有兴趣,[C(个人|不是优秀回答者)]	动态时间分布间隔低于均值,提问标题平均长度低于均值,上午,[高信息量]	0.27	0.78	2.53
积极→潜水	低影响力,其他地域,[C(个人|粉丝数低于均值|不是优秀回答者|回答者)]	低信息量	0.23	0.80	1.15
积极→积极	粉丝数低于均值,高创作等级,回答者,其他地域,[C(不是优秀回答者|个人)]	回答平均文本长度低于均值	0.30	0.82	1.14
积极→知识	高创作等级,高影响力,其他地域,未认证,[个人]	回答_消极	0.33	0.85	1.07
积极→需求	既是提问者又是回答者,粉丝数低于均值,[C(个人|不是优秀回答者)]	提问标题平均长度高于均值，上午，下午,[高信息量]	0.38	0.76	1.86
知识→潜水	粉丝数低于均值,个人,其他行业,[C(不是优秀回答者|回答者)]	回答平均文本长度低于均值	0.26	0.76	1.18
知识→积极	高影响力,回答者,关注数低于均值,个人,未认证,[不是优秀回答者]	高信息量,回答_消极	0.34	0.71	1.08
知识→知识	高创作等级,粉丝数高于均值,未认证,男性,不是优秀回答者,[C(个人|高影响力)]	高信息量,回答_消极	0.32	0.80	1.12
知识→需求	既是提问者又是回答者,男性,[C(个人|高影响力|未认证)]	回答_消极，回答平均文本长度高于均值,[C(晚上|高信息量)]	0.33	0.82	2.45
	既是提问者又是回答者,男性,[C(个人|高影响力|未认证)]	高信息量，深夜	0.33	0.82	2.45
需求→潜水	高创作等级，其他地域,未认证,[C(个人|回答者|不是优秀回答者)]	低信息量,动态时间分布间隔高于均值,[回答平均文本长度低于均值]	0.20	0.73	2.24
	高创作等级，其他地域,不是优秀回答者,[C(个人|回答者)]	低信息量,动态时间分布间隔高于均值,[回答平均文本长度低于均值]	0.20	0.73	2.24
需求→积极	粉丝数低于均值，回答者，其他地域,[C(个人|不是优秀回答者)]	动态时间分布间隔高于均值	0.26	0.77	2.14
需求→知识	关注数低于均值,回答者,[C(不是优秀回答者|高影响力)]	动态时间分布间隔高于均值，高信息量	0.35	0.75	1.59
	关注数高于均值,[C(个人|高影响力|不是优秀回答者)]	晚上，下午,[高信息量]	0.35	0.75	1.59
需求→需求	关注数高于均值,高创作等级,[C(个人|高影响力|未认证|不是优秀回答者)]	晚上,提问标题_积极,[C(提问标题平均长度高于均值|动态时间分布间隔低于均值|高信息量)]	0.46	0.80	1.49

的差异,潜水型和积极型向潜水型转变时,用户往往具有较低的影响力,前者倾向回答文本长度较低,后者倾向表现出低信息量;知识型用户向潜水型转变时,其个人粉丝数低于均值,这种转变可能因为某种契机让该用户在某一阶段成为了知识型用户,但却无法持续保持为知识型用户;需求型用户向潜水型转变时,该用户往往具有较高的创作等级,在行为上表现出了动态时间分布间隔高于均值和低信息量,说明用户虽为需求型用户但在吸收了一些知识后会回答部分问题,而其回答行为往往具有滞后性且回答的信息量并不高。

(2)当用户从不同角色向积极型角色转变的时候,潜水型和积极型用户向积极型转变时,表现出相同的行为模式,即回答文本长度较低,但前者往往具有低影响力,后者往往具有较高的创作等级;知识型用户和需求型用户的转变则表现出了完全不同的信息人特征和行为模式,知识型用户向积极型转变的时候,通常是高影响力用户回答文本且文本具有高信息量、回答情感比较消极,而需求型用户向积极型转变的时候,用户

粉丝数较少,在回答上直接表现出较高的回答滞后性。

(3)当用户从不同角色向知识型角色转变的时候,其文本具有高信息量特征是一个普遍的原则,但积极型用户向知识型用户转变是个例外,积极型向知识型转变的时候,用户具有高影响力和高创作等级,并会有 85% 的概率回答比较消极,回答文本的信息量于其而言不太重要;潜水型向知识型转变时,用户的粉丝数高于均值,并会有 70% 的概率回答出较长的文本内容且含有较高的信息量;当用户维持知识型角色不变的时候,用户本身往往具有高创作等级、粉丝数高于均值、高影响力,并会有 80% 的概率回答出具有高信息量和消极的内容;当用户从需求型向知识型转变的时候,当其关注数低于均值的时候,回答虽滞后但含有较高的信息量,当关注数高于均值的时候,偏爱在下午或者晚上的时候发布信息内容,信息内容也往往具有较高的信息量。

(4)当用户从不同角色向需求型角色转变的时候,其表现出了不同的信息发布时间倾向和文本长度倾向以及相同的高信息量文本特征。潜水型、积极型和知识型向需求型转变时,用户通常具有共同的特征,即既是提问者又是回答者。潜水型向需求型转变时,用户在对主题很有兴趣的情况下,会在上午提问或回答,且提问标题长度低于均值,回答时间间隔也低于均值;积极型向需求型转变的时候,粉丝数往往低于均值,并有 76% 的概率会在上午和下午提问或回答,且提问标题长度高于均值;知识型向需求型转变的时候,用户为高影响力,倾向于在深夜或者晚上提问或回答问题;当用户维持需求型不变的时候,用户关注数高于均值且具有高创作等级和高影响力,并且会有 80% 的概率喜欢在晚上提问,提问标题比较积极,提问标题的平均长度也高于均值。

4.6 基于主题的用户角色转变分析

如表 7 所示,为各角色转变条件下用户所关注的主题关联规则,因篇幅限制,仅展示提升度大于 2.5 的主题关联规则。基于主题的用户角色转变分析,以积极型用户向需求型用户转变为例,可以理解为用户从积极型向需求型转变时,用户最有可能同时关注群体免疫和美国死亡人数持续上涨两个话题,关注群体免疫的用户有 100% 的概率也会关注美国死亡人数持续上涨的话题。在所有角色转变条件下,用户所关注的主题基本围绕在美国引发的政治战争、美国医疗系统崩溃、群体免疫、比尔盖茨个人情况、美国死亡人数持续上涨、美国感染、确认及死亡人数等主题。在不同角色转变条件下,用户最有可能关注的话题也有区别。如潜水型用户向需求型用户转变的前提条件下,有 38% 的用户同时关注了美国医疗系统崩溃、西班牙流感和美国流感暴发、[C(群体免疫|美国死亡人数持续上涨)],如果用户主要关注“美国医疗系统崩溃”,在该转变条件下,用户有 100% 的可能性会对西班牙流感和美国流感暴发、[C(群体免疫|美国死亡人数持续上涨)]感兴趣;而需求型向需求型转变的条件下,用户则会同时关注冠状病毒抗体研究和新冠疫苗进入临床试验或者关注“散装江苏”支援各地并祈祷人民平安,希望渡过难关见阳光等,其他规则类似。通过对不同角色转变条件下用户所关注的主题进行分析,可以了解突发公共卫生事件下不同用户转变时的关注内容,对于平台及时推送相关问答具有一定意义,比如当用户维持潜水型角色不变的时候,可以观测潜水型用户向积极型转变的关注内容并通过计算潜水型和潜水型用户的相似性,推送类似的问答,刺激潜水型用户向积极型转变。

表 7 各角色转变条件下主题关联规则(部分)

<table>
<tr><th>角色转变</th><th colspan="2">关联规则(话题 = >话题)</th><th>支持度</th><th>置信度</th><th>提升度</th></tr>
<tr><td>潜水→积极</td><td>美国死亡人数持续上涨,美国引发政治战争,美国经济停摆</td><td>以反讽的语气为美国加油,比尔盖茨个人情况,新西兰防控策略</td><td>0.38</td><td>1</td><td>2.56</td></tr>
<tr><td>潜水→需求</td><td>美国医疗系统崩溃</td><td>西班牙流感和美国流感暴发,[C(群体免疫|美国死亡人数持续上涨)]</td><td>0.38</td><td>1</td><td>2.60</td></tr>
<tr><td>积极→积极</td><td>美国引发政治战争,新西兰防控策略</td><td>以反讽的语气为美国加油,比尔盖茨个人情况</td><td>0.38</td><td>1</td><td>2.59</td></tr>
<tr><td>积极→需求</td><td>群体免疫</td><td>美国死亡人数持续上涨</td><td>0.35</td><td>1</td><td>2.62</td></tr>
<tr><td>知识→知识</td><td>美国死亡人数持续上涨,美国经济停摆</td><td>以反讽的语气为美国加油,比尔盖茨个人情况,美国引发政治战争</td><td>0.38</td><td>1</td><td>2.54</td></tr>
<tr><td>需求→积极</td><td colspan="2">美国感染、确诊及死亡人数|美国医疗系统崩溃,美国死亡人数持续上涨</td><td>0.38</td><td>1</td><td>2.60</td></tr>
<tr><td>需求→需求</td><td colspan="2">散装江苏支援各地|祈祷人民平安,希望渡过难关见阳光</td><td>0.27</td><td>1</td><td>3.71</td></tr>
<tr><td></td><td colspan="2">冠状病毒抗体研究|新冠疫苗进入临床试验</td><td>0.27</td><td>1</td><td>3.71</td></tr>
</table>

5 结论与展望

本文以知乎平台中新型冠状病毒话题为例，首先从参与程度和价值维度，将问答平台不同生命周期阶段用户角色识别、划分为潜水型用户、积极型用户、需求型用户和知识型用户，然后构建基于信息生态理论的信息人、信息和信息环境的社区用户角色形成影响因素和最优解 Catboost 多分类器模型，并使用 SHAP 模型确定不同角色形成的关键影响要素，最后基于关联规则探讨用户角色转变的行为特点和主题特点。

本文的主要结论如下：①从全部转换阶段来看，用户倾向于维持角色不变且积极型维持不变的概率最高，为 49%，其他角色主要向积极型和潜水型用户转变。从不同转变阶段来看，以用户角色转换概率最大的潜水型 - 潜水型用户（即角色维持不变）为例，其不同转变阶段的特征变化程度具有显著差异，为维持角色不变，通常要求疫情发展后期粉丝数变化程度为上一阶段的 1.95 倍。②影响不同用户角色的关键因素不一致，但信息量始终是重要特征，信息量越低，反而越能说明该用户是潜水型用户。信息因子的时间属性是积极型用户的显著特征，说明积极型用户在投入社区的时候不在意时间段。信息环境对需求型和知识型用户形成的影响较大，其次是信息因子中的文本属性和情感属性。③用户向不同角色转变时，具有不同的转变规律和表现特征，如用户从积极型向需求型转变时，粉丝数往往低于均值，并以 76% 的概率在上午和下午提问或回答，且提问标题长度高于均值，用户也最有可能同时关注群体免疫和美国死亡人数持续上涨两个话题。

本研究的创新与贡献在于理论价值方面：①提供了突发传染病情境下用户细分模型和研究方法；②从信息生态视角构建用户角色形成的影响因素，并揭示影响不同角色形成的关键要素，拓展信息生态理论在社会化问答平台中的应用；③基于关联规则探究不同用户角色转变、行为模式和主题特点，是对用户角色研究的进一步深化。

研究结果对实践也有一定的启示：①对于问答社区，一方面，平台可以根据用户角色行为和主题特点提供个性化服务以促进用户向其他类型角色的良性转变，如向潜水型用户推荐与之具有相似主题的知识型用户，可以增加用户向积极型转变的概率；另一方面，平台可以依据用户角色形成的主要因素，针对性提升重要类型用户的占比，如研究显示良好的信息环境和主题内容，即热烈的讨论氛围和丰富的主题是需求型用户和知识型用户形成的关键。②对于政府而言，本研究所揭示 Covid-19 疫情期间用户角色在社区问答平台的行为和主题特点，有助于舆情管理部门了解用户转变的机制，尤其是不同角色类型转向需求型用户的行为和主题特点，通过及时满足信息需求来避免舆情恶化。

本研究还存在一定的局限性：①本文所研究的用户角色转变是涵盖了所有生命周期的角色转变情况，未具体细分和区分不同生命周期阶段的用户角色转变特点；②受限于知乎平台数据特征的获取，在衡量问题回答质量时使用指标较为单一，因信息疫情的存在可能导致研究存在一定的误差；③本文仅针对 Covid-19 话题下知乎平台的用户角色识别、形成和转变进行研究，研究结论对于其他事件情境的适用性和推广性还有待探索，后续还将继续研究其他传染病情境下更多问答平台上的用户角色形成与转变规律。另外基于本文的影响因素可以挖掘出更多的关联规则，但本文仅探讨行为和主题，未对其他关联因素进行分析。因此，在后续研究中，可以进一步对以上问题进行探索。

致谢：感谢图书情报国家级实验教学示范中心为本研究提供的实验支持！

参考文献：

[1] 人民网. 习近平在中央政治局常委会会议研究应对新型冠状病毒肺炎疫情工作时的讲话[EB/OL]. [2022 - 02 - 15]. http://cpc.people.com.cn/n1/2020/0215/c64094-31588554.html.

[2] 方陈承，张建同. 社会化问答社区中用户研究的述评与展望[J]. 情报杂志，2018，37(9)：185 - 193.

[3] 张薇薇，朱杰，蒋雪. 社会学习对专业虚拟社区不同类型用户知识贡献行为的影响研究[J]. 情报资料工作，2021，42(5)：94 - 103.

[4] ZHOU X, WU B, JIN Q. User role identification based on social behavior and networking analysis for information dissemination[J]. Future generation computer systems, 2019, 96: 639 - 648.

[5] 陈烨，王乐，陈天雨，等. 基于社会网络分析的社会化问答平台用户画像研究[J]. 情报学报，2021，40(4)：414 - 423.

[6] HAGEL J, ARMSTRONG A. Net gain: expanding markets through virtual communities[M]. Boston: Harvard Business Press, 1997.

[7] 张树森，梁循，齐金山. 社会网络角色识别方法综述[J]. 计算机学报，2017，40(3)：649 - 673.

[8] VILLODRE J, CRIADO J I. User roles for emergency management

in social media: understanding actors' behavior during the 2018 Majorca Island flash floods[J]. Government information quarterly, 2020, 37(4): 101521.
[9] PREECE J, SHNEIDERMAN B. The reader-to-leader framework: motivating technology-mediated social participation[J]. AIS transactions on human-computer interaction, 2009, 1(1): 13-32.
[10] FU C. Tracking user-role evolution via topic modeling in community question answering[J]. Information processing & management, 2019, 56(6): 102075.
[11] BARTAL A, RAVID G. Member behavior in dynamic online communities: role affiliation frequency model[J]. IEEE transactions on knowledge and data engineering, 2019, 32(9): 1773-1784.
[12] 赵欣,王倩雯,张长征. 从知识搜寻者到知识贡献者——专业虚拟社区用户角色转变的机理研究[J]. 情报科学,2017,35(10):18-22.
[13] ZENG G, GUAN H, CHEN F. Knowledge sharing in a virtual community of a hotel association: from free riders to active knowledge sharers[J]. Journal of China tourism research, 2014, 10(1): 95-119.
[14] ERYOMIN A L. Information ecology-a viewpoint[J]. International journal of environmental studies, 1998, 54(3/4): 241-253.
[15] 马海群,李钟隽,张涛. 数据新闻信息生态链形成、演替及运行价值分析[J]. 科技情报研究,2021,3(3):49-59.
[16] NARDI B A, O'DAY V, O'DAY V L. Information ecologies: using technology with heart[M]. Cambridge: MIT Press, 1999.
[17] 娄策群,赵桂芹. 信息生态平衡及其在构建和谐社会中的作用[J]. 情报科学,2006(11):1606-1610.
[18] 张海涛,耷毓琦,刘阔,等. 生态信息:内涵、特点、运动过程与运动规律研究[J]. 图书情报工作,2013,57(12):46-50.
[19] 窦悦. 信息生态视角下"3×3"应急情报体系构建研究[J]. 图书情报工作,2020,64(15):82-89.
[20] FINK S. Crisis management: planning for the Inevitable[M]. New York: American Management Association, 1986: 20.
[21] BURKHOLDER B T, TOOLE M J. Evolution of complex disasters[J]. The lancet, 1995, 346(8981): 1012-1015.
[22] 安璐,陈苗苗,李纲. 社交媒体环境下突发事件严重性评估和预警机制研究[J]. 图书情报工作,2021,65(5):98-109.
[23] 刘冰,肖高飞,晁世育. 重大突发公共卫生事件风险研判与决策模型构建研究[J]. 信息资源管理学报, 2021,11(5):17-26,37.
[24] 姜金贵,闫思琦. 基于主题和情绪相互作用的微博舆情演化研究——以"红黄蓝虐童事件"为例[J]. 情报杂志,2018,37(12):118-123.
[25] 安璐,吴林. 融合主题与情感特征的突发事件微博舆情演化分析[J]. 图书情报工作,2017,61 (15):120-129.
[26] 陈娟,高杉,邓胜利. 社会化问答用户特征识别与行为动机分析——以"知乎"为例[J]. 情报科学,2017,35(5):69-74,80.
[27] 韩世曦,曾粤亮. 突发公共卫生事件背景下数字青年微信公众平台健康信息采纳意愿影响因素研究[J]. 图书馆学研究, 2021(6):83-92.
[28] 李胜利,钟滢. 中外技术问答社区的实证对比研究与启示——以 CSDN 和 Stack Overflow 为例[J]. 情报学报,2020,39(9): 989-1000.
[29] 韩盟,吴红,李昌,等. 高校可转移专利识别研究——基于贝叶斯理论和组合赋权法[J]. 图书情报工作,2021,65(5):118-125.
[30] RIDINGS C, GEFEN D, ARINZE B. Psychological barriers: lurker and poster motivation and behavior in online communities[J]. Communications of the Association for Information Systems, 2006, 18(16):329-354.
[31] SHANNON C E. A mathematical theory of communication[J]. Bell system technical journal, 1948,27:370-423.
[32] 赵丹,王晰巍,李师萌,等. 新媒体环境下的网络舆情特征量及行为规律研究——基于信息生态理论[J]. 情报学报,2017,36(12):1224-1232.
[33] 李嘉兴,王晰巍,李师萌,等. 信息生态视角下老年用户群体微信使用行为影响因素研究[J]. 图书情报工作,2017,61(15): 25-33. DOI:10.13266/j.issn.0252-3116.2017.15.003.
[34] TIAN H, GAO C, XIAO X, et al. SKEP: sentiment knowledge enhanced pre-training for sentiment analysis[J]. ArXiv preprint arXiv:2005.05635, 2020.
[35] Github. BERTopic[EB/OL]. [2021-12-04]. https://github.com/MaartenGr/BERTopic.
[36] Github. SHAP[EB/OL]. [2022-03-15]. https://github.com/slundberg/shap.
[37] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[C]//International conference on intelligent computing. Berlin: Springer, 2005: 878-887.
[38] BORGELT C. An implementation of the FP-growth algorithm[C]//Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations. New York: Association for Computing Machinery, 2005: 1-5.

作者贡献说明:
陈苗苗:提出研究思路,负责文献搜集、研究方案设计与数据收集处理,撰写及修改论文;
安璐:指导研究思路和研究方法,提出修改意见,修改论文。

陈苗苗，安璐. 突发传染病情境下社会化问答平台用户角色形成及转变——以知乎平台为例[J]. 图书情报工作,2022,66(12):68 - 81.

User Role Formation and Transformation of Socialized Q&A Platforms in the Context of Infectious Disease Outbreaks: Taking the Zhihu Platform as an Example

Chen Miaomiao[1] An Lu[1,2]

[1] School of Information Management, Wuhan University, Wuhan 430072

[2] Center for Studies of Information Resources, Wuhan University, Wuhan 430072

Abstract: [**Purpose/Significance**] To explore the user role classification methods, key factors of role formation, transformation characteristics and differences of the Q&A platforms in the context of infectious disease outbreaks. [**Method/Process**] A total of 702,927 data related to Covid-19 epidemic were collected from Q&A platforms. The user roles were analyzed from the dimensions of participation and value. The influencing factors of community user role formation were constructed based on the information user factor, information factor and information environment factor. The key factors affecting the formation of different roles were analyzed by combining the multi-classification model and the SHapley Additive exPlanations (SHAP) model. The FP-growth association rule algorithm was used to mine behavior patterns and topic characteristics during the transformation of different roles. [**Result/Conclusion**] The results show that users tend to keep their roles unchanged, and the transformation direction is mainly towards active or diving roles. The amount of information is the key factor for the formation of different roles. There are significant differences in the extent of change in user role transformation characteristics in different transformation stages and user role transformation behaviors in all transformation stages.

Keywords: user role Zhihu Q&A platform role transformation influencing factors Covid-19 infectious disease outbreak